DOI:
10.1016/j.jad.2015.07.015

Document Version
Peer reviewed version

Post-print version

http://www.jad-journal.com/article/S0165-0327(15)30047-1/abstract

Hanlon, C., Medhin, G., Selamu, M., Breuer, E., Worku, B., Hailemariam, M., Lund, C., Prince, M. and Fekadu, A. Validity of brief screening questionnaires to detect depression in primary care in Ethiopia (2015). Journal of Affective Disorders. 186: 32-39.

Validity of brief screening questionnaires to detect depression in primary care in Ethiopia

Charlotte Hanlon[*a,b], Girmay Medhin[c], Medhin Selamu[a], Erica Breuer[d], Benyam Worku[a], Maji Hailemariam[a], Crick Lund[d], Martin Prince[b] and Abebaw Fekadu[a,e]

[*]Corresponding author

[a]Addis Ababa University, College of Health Sciences, School of Medicine, Department of Psychiatry

[b]King's College London, Institute of Psychiatry, Psychology and Neuroscience, Health Services and Population Research Department, Centre for Global Mental Health

[c]Addis Ababa University, Aklilu-Lemma Institute of Pathobiology

[d]Alan J Flisher Centre for Public Mental Health, Department of Psychiatry and Mental Health, University of Cape Town

[e]King's College London, Centre for Affective Disorders, Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neuroscience

[*] Department of Psychiatry, 6th Floor of the College of Health Sciences Building, Tikur Anbessa Hospital, Addis Ababa, PO 9086, Ethiopia. Telephone: +251 912 803374. Email: charlotte.hanlon@kcl.ac.uk.

Abstract

Background

Brief depression screening questionnaires may increase detection of depression in primary care settings but there have been few validation studies carried out in typical populations in low-income countries.

Methods

Cultural validation of the Patient Health Questionnaire (PHQ-9/ PHQ-2), the 20-item Self-Reporting Questionnaire (SRQ-20) and the Kessler scales (K6/ K10) was carried out in 306 adults consecutively attending primary care facilities in small towns in Ethiopia. To assess criterion validity, the gold standard assessment for presence of Major Depressive Disorder (MDD) was made by Ethiopian psychiatric nurses using the Mini International Neuropsychiatric Interview.

Results

The prevalence of gold standard MDD was 5.9%, with irritability more common than depressed mood or anhedonia. The area under the receiver operating characteristic curve indicated good performance of the PHQ-9, SRQ-20, K6 and K10 (0.83 to 0.85) but only fair for the PHQ-2 (0.78). No cut-off score had acceptable sensitivity combined with adequate positive predictive value. All screening questionnaires were associated with disability and the PHQ-9 and SRQ-20 were associated with higher health service contacts, indicating convergent validity. Construct validity of all scales was indicated by unidimensionality on exploratory factor analysis.

Limitations

Test-retest reliability was not assessed.

Conclusions

Brief depression screening questionnaires were found to be valid in primary care in this low-income country. However, these questionnaires do not have immediate applicability in routine clinical settings. Further studies should evaluate utility of indicated screening embedded within health system changes that support MDD detection. Investigation of irritability as a core depression symptom is warranted.

Keywords:

Depression; validation; primary health care, sub-Saharan Africa; developing country; screening

1. Introduction

The treatment gap for depression in low- and middle-income countries (LMICs) is large; in the World Mental Health Survey in China and Nigeria, fewer than 10% of people with depression had received treatment in the year of illness onset [1]. Undetected depression in rural low-income country community settings has been associated with substantial disability, premature mortality and increased attendance at health facilities with somatic complaints [2]. Non-detection of depression in primary healthcare (PHC) settings is associated with cost and time burdens to both patient and health system due to repeated attendance, inappropriate prescribing of non-evidence-based interventions and poorer prognosis of co-morbid physical health conditions [3].

Effective treatments for depression that are appropriate and feasible for LMICs exist [4, 5] but are not available to the majority of the population [6]. In order to reduce the treatment gap for depression, the World Health Organization's 'mental health gap action programme' (mhGAP) advocates integration of mental health care into PHC services through task sharing, supported by new evidence-based treatment guidelines suitable for PHC workers [7, 8]. Depression is a priority disorder for mhGAP; however, for successful implementation, increased community awareness and demand for care needs to be coupled with improved detection of depression within PHC.

In high-income countries, accurate detection of depression in PHC is more likely in the presence of a good therapeutic relationship, greater clinical experience, frequent patient contacts and when the patient presents with psychological, rather than somatic, symptoms of depression [9]. This presents a great challenge in LMICs where the PHC system is often weak, with over-burdened clinics and high turnover of staff who have

relatively limited training and clinical experience. In addition, somatic presentations of depression are the norm in this setting [10].

Improving detection of depression by PHC workers is challenging [11]. In high-income countries, stand-alone training is ineffective and the use of brief, screening questionnaires for depression are only effective in improving recognition, treatment and outcomes of depression if applied in conjunction with organisational changes, such as collaborative care [12, 13]. However, in low-income countries brief depression screening tools may have greater utility.

In this paper we present the cultural validation of brief questionnaires for detecting depression in PHC settings in Ethiopia. The objective was to examine the potential application of such questionnaires to the implementation and evaluation of mental health care integrated into primary care in Ethiopia as part of the Programme for Improving Mental health CarE (PRIME) [14].

2. Material and methods

2.1 Study design

A validation study to investigate the criterion, convergent and construct validity of brief depression screening questionnaires.

2.2 Setting

The study was carried out in health centres located in and around Butajira town, 130km south of the capital city, Addis Ababa, Ethiopia. A psychiatric nurse-led unit is

located in the Zonal hospital in Butajira town. However, at the time of the study there was no mental health care available in PHC. The PHC facility is staffed by health officers and nurses. This validation study was carried out in one urban, two semi-urban and one rural health centre, selected purposively.

2.3 Brief questionnaires to detect depression

(1) Self Reporting Questionnaire, 20-item version (SRQ-20) [15]

The SRQ-20 was developed to improve detection of common mental disorders in PHC settings in LMICs [15]. The SRQ items are single-clause questions which require a yes/no response and are easily administered in an interview format. The items include somatic, anxiety and depressive symptoms, as well as questions about suicidal ideation and functional impact, present in the preceding 30 days. The SRQ-20 has been widely validated and used in LMICs. In Ethiopia, the SRQ-20 has been validated for detection of depression in postnatal women [16, 17]. A culturally adapted version of SRQ ('SRQ-F') lengthened the scale with little improvement in psychometric properties [18].

(2) Kessler 6 and 10 item versions (K6/ K10) [19]

The K10 and K6 scales are widely used tools to assess non-specific psychological distress in the previous one month. Each item is rated from 1 to 5, based on the persistence of a specific symptom, from "none at all" to "all the time". The K10 scale includes depressive, anxiety and somatic symptoms but not suicidal ideation. The Kessler 6-item scale comprises a sub-set of the Kessler 10-item scale (items 2, 3, 5, 7, 9 and 10). In Ethiopia, the criterion validity of K6 and K10 for detection of depression

in postnatal women has been demonstrated [17] but the Kessler has not been evaluated in a general PHC setting.

(3) Patient Health Questionnaire, 2- and 9-item versions (PHQ-2 [20] and PHQ-9 [21])

The PHQ-9 is a widely used depression screening scale for PHC in high-income countries. The nine items of the PHQ follow the Diagnostic and Statistical Manual (DSM) version IV [22] diagnostic criteria for a depressive episode, including suicidal ideation, and ask about symptoms present in the preceding two weeks. The four response categories refer to the amount of time that the symptom was present (from 'not at all' (0) to 'nearly every day' (3)). A tenth item asks about the functional impact of the symptoms. Responses to the PHQ-9 can be summed to give a total symptom score. Alternatively, the DSM diagnostic algorithm for a major depressive episode can be applied to give a categorical diagnosis of depression. The PHQ-2 is comprised of the first two items of the PHQ-9 and has been found to be a useful screener in PHC settings in high-income countries [20]. The criterion validity of the PHQ-9 for detecting depression has been demonstrated in medical out-patients at a referral hospital in Addis Ababa, Ethiopia [23].

2.4 Semantic, technical and content validation

Semantic, technical and content validity in the Ethiopia setting have already been established for the Amharic versions of the SRQ-20 [16, 24] and K10 [17]. The PHQ-9 was translated independently into Amharic by two Ethiopian psychiatrists and then back-translated into English. The final version was obtained by expert consensus. Moving from a self-completed to an interviewer-administered questionnaire

necessitated modifications, in keeping with previous studies from low-income countries [25]. See Supplementary Files 1 and 2.

2.5 Criterion, construct and convergent validation study

Sample

Consecutive patients attending the four PHC facilities were recruited over a three week period in March 2013. Patients were approached after they had consulted the PHC worker. Patients were excluded if they required emergency medical attention, were unable to converse in Amharic (the official language of Ethiopia), were unable to communicate or were suffering from severe mental disorder.

Procedure

Participants were interviewed by (1) lay data collectors who administered the brief depression scales (SRQ-20, K10/K6 and PHQ-9/PHQ-2) and other structured questionnaires (see below), and (2) psychiatric nurses who carried out a gold standard clinical assessment to determine the presence or absence of MDD. To overcome order effects, the administration of the scales and assessments was randomised. Psychiatric nurses and lay data collectors were each masked to the outcome of the other assessment.

Gold standard measure of depression

The MINI is a structured diagnostic assessment scale that allows DSM-IV and ICD-10 diagnosis [26] and has been widely used as a gold standard measure of MDD [27]. Adaptations to the MINI MDD module were made in order to improve sensitivity to sociocultural expressions of distress. The skip rules were not applied and raters were

allowed to explore symptoms once the fully-structured probe questions had been asked. A question on irritability was added, based on previous qualitative work and clinical experience that indicate this is an important manifestation of depression in the setting. After administering the MINI, combined with information gleaned from further clinical interview and observation, and considering bereavement and organic exclusion criteria, the psychiatric nurses gave their overall clinical judgment as to whether the person was suffering from MDD.

Other measures

The convergent validity of the brief depression scales was evaluated by examining the association with frequency of health service use in the preceding three months and functional impairment. The World Health Organization Disability Assessment Schedule, version 2 (WHODAS 2.0) [28] assesses both the level of functional impairment and the number of days lost from work in the previous 30 days due to a health condition. The instrument is considered to be cross-culturally applicable and the Amharic version has been used in Ethiopia previously [29].

Sociodemographic characteristics (age, sex, marital status, educational level, area of residence and occupation) were measured using structured self-report questions.. Participants were also asked systematically about their main reasons for seeking help from the health centre and the diagnosis that the health worker had given them.

Training

The lay-interviewers were all high school completers who are experienced data collectors in the field of mental health. The lay data collectors were trained for four

days by senior mental health researchers, followed by one day of observed pilot interviews.

Three psychiatric nurses were trained by an Ethiopian psychiatrist for two days in administration of the criterion measure (MINI). The training included extensive role play and observed piloting of the gold standard assessment.

Sample size

A formula for calculating sensitivity and specificity for single tests was used [30]. The assumptions used for calculating sample size were as follows: anticipated sensitivity 80%, alpha 0.05, L (desired precision) = 0.1, $Z^2_{1-\alpha/2}$=3.84, prevalence = 20% and loss due to incomplete data or other reasons=3%. The required sample size was 316, to allow calculation of the sensitivity estimate within the confidence limits of 75% and 84%.

Data management and quality assurance

Data were checked for completeness in the field by supervisors and entered on the day of data collection where possible. Double data entry with consistency checks was carried out using EpiData [31]. An Ethiopian psychiatrist supervised the psychiatric nurse assessments. An experienced supervisor observed selected lay interviewer assessments.

Data analysis

Data analysis was carried out using Stata version 13.1 [32]. Receiver Operating Characteristic (ROC) curves were plotted. Optimal cut-off scores were identified on the basis of the maximum specificity which was not higher than sensitivity. Internal

consistency was evaluated using Cronbach's alpha [33]. At the optimal cut-point, sensitivity, specificity, positive and negative predictive values, percentage correct classification, Cohen's kappa and Youden's index (sensitivity + specificity -1) [34] were calculated.

The prevalence ratio for individual depression screening scale items in people with MDD vs. people without MDD was calculated using a poisson working model and sandwich estimators of the standard errors [35]. Items on the PHQ-9 were dichotomised as 0/1 vs. 2/3, and items on the Kessler were dichotomised as 1/3 vs. 4/5.

Convergent validity was evaluated using non-parametric tests. Spearman's Rho correlation coefficient was calculated for the association between total score on the WHODAS 2.0/number of days when unable to function and the total scores of each depression screening scale. Kruskal-Wallis test of the equality of medians was used to compare the distribution of depression screening scale scores in people who had attended the health centre in the previous three months on no occasions, once or two or more times.

Construct validity of the depression screening scales was investigated using exploratory factor analysis with maximum likelihood extraction and varimax rotation. Inspection of eigenvalues and scree plots was used to identify the number of factors to be retained as indicators of scale dimensions. As the SRQ-20 has dichotomous response categories, tetrachoric correlation matrices were used as the basis for factor analysis.

In order to investigate misclassification of MDD by the screening scales, a multivariable logistic regression analysis was carried out with each scale score dichotomised at the optimal cut-off score for detection as the dependent variable. By including the gold standard diagnosis of MDD as an independent variable, all other associations with the dependent variable were indicative of misclassification. The following items were identified as being potentially associated with misclassification of MDD: age, sex [36], education, rural vs. urban residence and two items on the WHODAS 2.0 ("difficulty standing for long periods" and "difficulty walking long distances") that were more indicative of physical than mental disability [37].

Ethical considerations

The investigation was carried out in accordance with the latest version of the Declaration of Helsinki, the study design was reviewed by the Institutional Review Board of the College of Health Sciences, Addis Ababa University and informed consent of the participants was obtained after the nature of the procedures had been fully explained. All persons identified by the psychiatric nurses as suffering from mental disorder were offered appropriate treatment and follow-up with the existing Butajira psychiatric service.

Results

A total of 309 adults were recruited into the study. There were no refusers. Data were missing for three participants, giving a final sample of 306. The socio-demographic characteristics of participants are described in Table 1.

[Table 1]

3.1 Gold standard MDD

The prevalence of major depressive disorder was 5.9% (n=18) according to the gold standard psychiatric nurse assessment using the MINI. Only 11 out of 18 people with MDD had either of the core DSM-IV depression symptoms (depression or anhedonia); however a further six cases of MDD had mood change in the form of irritability (*"being easily angry or upset"*). Overall, 14/18 cases of MDD had irritability (19/286 for non-cases), 9/18 of cases had depressed mood (4/284 non-cases) and 8/18 of cases had anhedonia (7/287 non-cases).

None of the participants with MDD reported any emotional symptoms as being their primary reason for seeking help; however, sleep and appetite problems and 'anxiety' and 'depressed mood' were reported as secondary reasons for attendance (n=3). Two participants were undergoing voluntary testing and counselling for HIV, one was attending antenatal care and the remainder presented with abdominal complaints (n=5), headache (n=4), burning sensations, palpitations, cough, and toothache (n=1 each). The majority of participants reported that the PHC workers had not informed them of the diagnosis. None had been given a diagnosis of MDD by the PHC worker.

3.2 Depression screening scales

The frequency distributions for all of the depression scales were negatively skewed. The median scores (25^{th} and 75^{th} centiles) were as follows: PHQ-9: 12 (10, 18), PHQ-2: 6 (6, 10), SRQ-20: 3 (1, 8), K10: 12 (10, 18) and K6: 6 (6, 10). There was no evidence of ceiling effects, but possible floor effects for the PHQ-2 (58.2% of participants scored 0 out of 6) and Kessler scales (52.3% scored the minimum score

on the K6 and 42.2% for the K10). Internal consistency, as indicated by Cronbach's alpha, was good (0.84 for PHQ-9, 0.90 for SRQ-20, 0.88 for K10 and 0.85 for K6).

3.3 Criterion validity

See Table 2. The area under the receiver operating characteristic curve of the depression screening scales against the gold standard was very similar for the PHQ-9 (0.85; 95% confidence interval 0.77, 0.95), SRQ-20 (0.84; 95%CI 0.76, 0.91), K10 (0.83; 95%CI 0.74, 0.92) and K6 (0.84; 95%CI 0.75, 0.92), but lower for the PHQ-2 (0.78; 95%CI 0.66, 0.90). See Figure 1.

[Figure 1]

Apart from the PHQ-2, the optimal scale cut-off scores for indicating probable MDD for the other scales had Youden's scores > 0.50 and correct classification in around three-quarters of cases; however, the positive predictive value at these cut-off scores was very low: under 18% for all scales and only 11.7% for the PHQ-2. See Supplementary Table 1.

[Table 2]

When the cut-off score was increased so that the PPV was at least 50%, as per recommendations for useful depression scales in PHC settings [38], this led to an unacceptable drop in sensitivity. The agreement between the brief screening scales and the gold standard was only moderate, as indicated by the kappa scores which clustered around 0.20.

Applying the DSM-IV diagnostic algorithm to the PHQ item responses, 2.9% (n=9) fulfilled criteria for MDD. The PHQ algorithmic diagnosis of MDD had very low sensitivity (22.0%) and low PPV (44.0%).

Most of the individual items on each of the depression screening scales were significantly more prevalent in people with MDD compared to those without MDD. See Table 3.

[Table 3]

The prevalence ratios for items in the PHQ-9 were higher than for most of the other scales, in particular when compared to the SRQ-20. Somatic symptoms tended to discriminate less well than cognitive symptoms between people with and without MDD.

3.4 Convergent validity

[Table 4]

All of the depression screening scales were highly correlated with WHODAS 2.0 disability score and the number of days of disability in the preceding month (See Table 4). Increasing numbers of previous visits with a health centre were associated with higher scores on the PHQ-9 and the SRQ-20 but not on the PHQ-2 or the Kessler scales.

3.5 Construct validity

The PHQ-9 and Kessler 10 were unidimensional on exploratory factor analysis. For these scales, all items loaded onto the resulting factor with an item-factor correlation of 0.35 or more. For the SRQ-20, two factors had Eigenvalues greater than 1.0 but there was substantial cross-loading of items and, therefore, a single factor solution was considered most meaningful. Items 16 (worthlessness) and 18 (fatigue) of the SRQ-20 did not load onto any of the factor models. See Supplementary Figure 1 and Supplementary Tables 2 to 4.

3.6 Misclassification

See Table 5. In the multivariable model, misclassification of MDD was associated with lower educational level and the presence of physical disability.

[Table 5]

Discussion

In this validation study of brief depression screening questionnaires in PHC settings in Ethiopia, there was evidence of criterion, convergent and construct validity for the PHQ-9, SRQ-20, K10 and K6. The two item version of the PHQ (PHQ-2) had less good criterion validity and the PHQ DSM-IV diagnosis of MDD was extremely insensitive. All scales were affected by measurement bias, with lower educational level and the presence of physical disability associated with misclassification of MDD. None of the scales combined sufficiently high positive predictive values and

sensitivity for routine screening for MDD in unselected populations in clinical settings.

The similarity in the psychometric properties of the PHQ-9, SRQ-20, K6 and K10 for detection of MDD is perhaps surprising, although in keeping with a previous head-to-head comparison in India [38]. The PHQ-9 and K10/6 item response categories measure the persistence of symptoms whereas the SRQ-20 item response categories only measure the existence of symptoms, regardless of severity or persistence. In the low-literacy setting of this study it is possible that the potential gains in accuracy were offset by the complexity of the multiple response categories. The reporting time periods also differ between the scales: 30 days for the SRQ-20 and K10/6, and two weeks for the PHQ-9,but this did not result in a higher prevalence of 'probable MDD' at the optimal cut-off point. Furthermore, the SRQ-20 was designed as a measure of common mental disorders, combining depressive, anxiety and somatic symptoms, and the K10/6 similarly measures non-specific psychological distress whereas the PHQ-9 items closely follow the DSM-IV diagnostic criteria. This finding accords with our previous validation work in Ethiopia [39], whereby the SRQ-20 was found to be a more valid screening measure of MDD in the postnatal period than a specific depression measure.

The dramatic loss in sensitivity of the PHQ-9 resulting from application of the DSM-IV diagnostic criteria raises concerns about the assumption that the core symptoms of MDD (depressed mood and anhedonia) are the same across cultures. In our study, only 11 out of 18 participants diagnosed as having MDD by the experienced Ethiopian mental health professionals reported either of the core mood symptoms.

Seven out of the other eight people with MDD presented with irritability, which was added to the MINI for our study due to the perceived usefulness in clinical practice in Ethiopia. Irritability is included in DSM-V as a core symptom of depression in young people [40]. In the National Co-morbidity Replication Survey from the US, irritability was a common symptom in people with MDD, although was rarely present in the absence of depressed mood [41]. The addition of irritability to core diagnostic criteria for depression may be more applicable within Ethiopia. Although the area is complex and requires further study, irritability and anger are strongly proscribed in Ethiopian culture and identified readily as deviant states. The same may not be true for sadness, which may provoke less societal reaction. As we have argued previously, the applicability of the concept of anhedonia in a collectivist and religious culture may be limited and contribute to low reporting of this symptom [16].

The optimal cut-off scores for detecting MDD with the brief questionnaires were lower than those found in validation studies conducted in high-income countries. A recent review of optimal cut-off scores for the PHQ-9 [42] found that PHQ-9 cut-off scores ranging between 8 and 11 performed well in identifying probable MDD. That review included five studies from middle-income countries but no low-income countries. However, studies of PHQ-9 validity in low-income countries from sub-Saharan Africa (Cameroon and Ghana) found low sensitivity at the standard cut-off point [37, 43]. African validation studies of the PHQ-9 which have found higher cut-off points were conducted in well-educated, urban populations [44, 45].

The performance of the brief depression questionnaires was limited by measurement bias in this study, with misclassification associated significantly with lower

educational level and physical disability. Previously we found that depression screening questionnaires that were developed in Western country settings perform better in urban than rural Ethiopia [16, 17]. Mental health literacy, which is related to general levels of education, is likely to play an important part in the performance of self-report depression questionnaires. Bias due to misclassification of symptoms of physical illness as expressions of emotional distress is a recognised problem in applying depression screening scales in a PHC setting with high burden of physical health problems. In this study, somatic symptoms were less discriminating than items that were more cognitive in nature. However, the most frequently endorsed symptoms in people with gold standard MDD were somatic in nature. Furthermore, the symptoms motivating attendance at the PHC facility were all somatic in nature, underlining the importance of somatic symptoms in detection of depression in the PHC setting.

A limitation of the study was that we were unable to evaluate test-retest reliability due to feasibility constraints.

Conclusions

Overall there was little to distinguish between the validity of the PHQ-9, SRQ-20 and K10/6. The PHQ-9 and SRQ-20 had better convergent validity when compared to the K10/6 as well as the advantage of incorporating routine enquiry about suicidal ideation. The theoretical advantage of the PHQ-9 in allowing diagnosis of MDD using DSM criteria was not realised in practice and raises questions about the direct applicability of diagnostic algorithms compared to simple symptom counts. The

utility of ‘irritability’ as an important mood manifestation of MDD in this Ethiopian setting requires further investigation.

Figure 1: Receiver Operating Characteristics Curves for the Patient Health Questionnaire (2- and 9-item), Self-Reporting Questionnaire and Kessler scales (6- and 10-item)

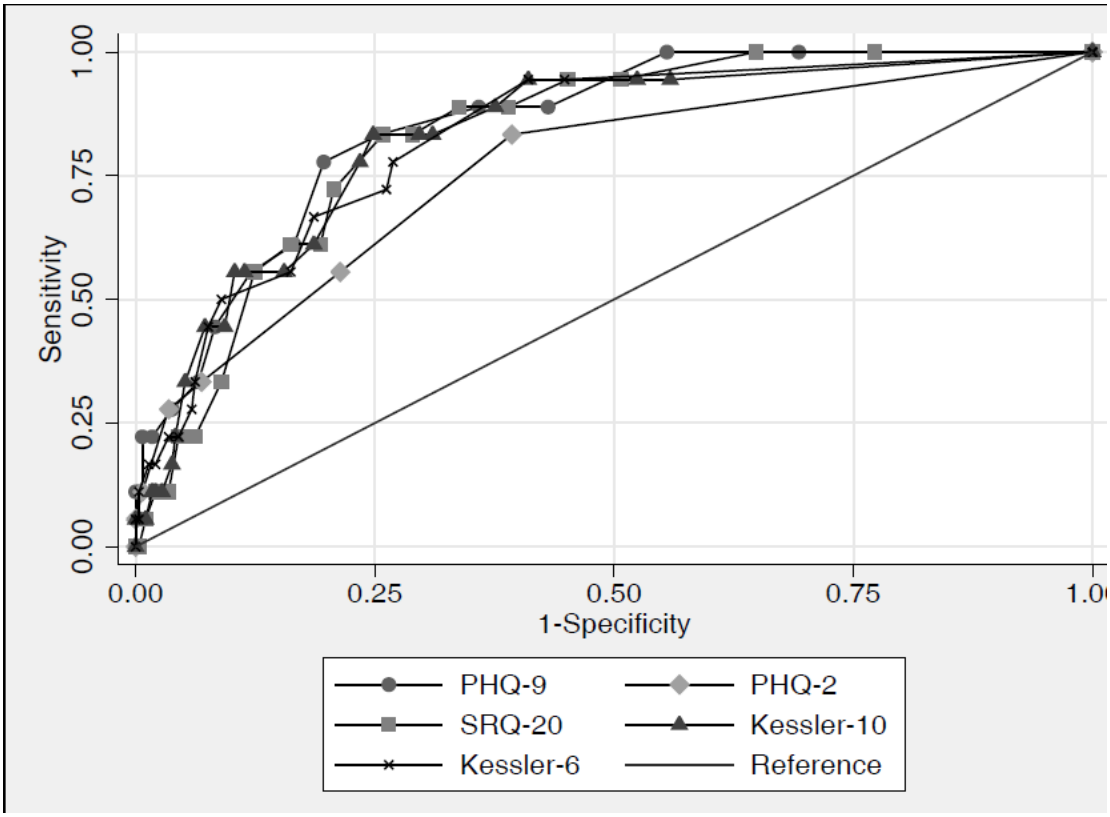

Table 1: Characteristics of validation sample attending primary health care centre (n=306)

Characteristic		*N (%)*
Sex		
	Male	115 (37.6)
	Female	191 (62.4)
Age (years)		
	< 20	47 (15.4)
	20 to 29	137 (44.8)
	30 to 39	62 (20.3)
	40 to 49	28 (9.2)
	50 or more	32 (10.4)
Marital status		
	Never married	112 (37.2)
	Married	176 (58.5)
	Widowed or divorced	13 (4.3)
Educational status		
	No education	69 (22.7)
	Primary	62 (20.4)
	Secondary	85 (28.0)
	11^{th} grade or above	88 (29.0)
Occupation		
	Housewife	74 (24.2)
	Student	55 (18.0)
	Government employee	48 (15.7)
	Farmer	40 (13.1)
	Merchant	33 (10.8)
	Daily labourer	33 (10.8)
	Unemployed	12 (3.9)
	Other	11 (3.6)
Residence		
	Urban	193 (63.3)
	Rural	112 (36.7)

Table 2: Optimal cut-off scores for detection of Major Depressive Disorder (MDD) and associated validity coefficients

	Prevalence of probable MDD	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	% classified correctly	Youden's index	Kappa
Patient Health Questionnaire (PHQ)-9								
≥4	38.9	88.9	64.2	13.5	98.9	65.7	0.53	0.15
≥5	**28.8**	**83.3**	**74.7**	**17.1**	**98.6**	**75.2**	**0.58**	**0.21**
≥6	22.9	77.8	80.6	20.0	98.3	80.4	0.58	0.25
PHQ-2								
≥1	**41.8**	**83.3**	**60.8**	**11.7**	**98.3**	**62.1**	**0.44**	**0.11**
≥2	23.2	55.6	78.8	14.1	96.6	77.5	0.34	0.14
≥3	8.5	33.3	93.1	23.1	95.7	89.5	0.26	0.22
PHQ diagnosis of MDD	2.9	22.0	98.0	44.0	95.0	94.0	0.20	0.27
Self-Reporting Questionnaire-20								
≥7	32.0	83.3	71.2	15.3	98.6	71.9	0.55	0.18
≥8	**29.1**	**83.3**	**74.3**	**16.9**	**98.6**	**74.8**	**0.58**	**0.20**
≥9	23.5	72.2	79.5	18.1	97.9	79.1	0.52	0.22
Kessler-10								
≥17	28.1	83.3	75.4	17.4	98.6	75.8	0.59	0.21
≥18	**26.5**	**77.8**	**76.7**	**17.3**	**98.2**	**76.8**	**0.55**	**0.21**
≥19	20.9	61.1	81.6	17.2	97.1	80.4	0.43	0.19
Kessler-6								
≥8	44.1	94.4	59.0	12.6	99.4	61.1	0.53	0.13
≥9	**29.7**	**77.8**	**73.3**	**15.4**	**98.1**	**73.5**	**0.51**	**0.18**
≥10	28.8	72.2	74.0	14.8	97.7	73.9	0.46	0.16

Table 3: Discrimination between cases and non-cases of major depressive disorder (MDD) by items of brief depression screening scales

	Prevalence in non-cases (%)	Prevalence in people with gold standard MDD (%)	Prevalence ratio (95% confidence interval)
Patient Health Questionnaire, 9 item			
Trouble concentrating	0.7	11.1	16.00 (2.38, 107.44)
Feeling down, depressed or hopeless	3.1	27.8	8.89 (3.32, 23.82)
Feeling bad about yourself	2.8	16.7	6.00 (1.74, 20.74)
Retardation or agitation	2.8	16.7	6.00 (1.74, 20.74)
Tired or low energy	8.6	50.0	5.76 (3.17, 10.45)
Appetite problem	9.3	50.0	5.33 (2.97, 9.58)
Sleep problem	8.6	44.4	5.12 (2.70, 9.70)
Little interest or pleasure	9.0	38.9	4.31 (2.17, 8.56)
Better off dead	9.0	22.2	2.46 (0.96, 6.30)
Self Reporting Questionnaire, 20-item			
Thought of ending life	5.2	22.2	4.25 (1.57, 11.52)
Difficult to make decisions	9.7	38.9	4.13 (2.09, 8.18)
Feel worthless	13.8	50.0	3.68 (2.14, 6.36)
Trouble thinking clearly	21.0	72.2	3.47 (2.41, 4.99)
Feel unhappy	21.7	66.7	3.10 (2.09, 4.60)
Poor digestion	26.9	83.3	3.08 (2.32, 4.07)
Hands shake	10.7	27.8	2.67 (1.17, 6.05)
Difficult to enjoy activities	12.8	33.3	2.59 (1.26, 5.31)
Easily frightened	24.5	61.1	2.51 (1.65, 3.83)
Lost interest in things	19.4	44.4	2.32 (1.31, 4.10)
Feel tired	29.4	66.7	2.28 (1.57, 3.31)
Easily tired	34.3	77.8	2.28 (1.70, 3.06)
Daily work suffering	24.6	55.6	2.28 (1.44, 3.61)
Unable to play useful part in life	18.3	38.9	2.15 (1.14, 4.03)
Nervous, tense or worried	34.1	72.2	2.12 (1.53, 2.95)
Sleep badly	32.1	66.7	2.09 (1.44, 3.02)
Crying more than usual	8.6	16.7	2.00 (0.66, 6.03)
Poor appetite	39.3	77.8	1.98 (1.49, 2.64)
Uncomfortable stomach	25.3	44.4	1.77 (1.02, 3.09)
Frequent headaches	49.7	72.2	1.45 (1.07, 1.98)
Kessler 10 item (Kessler 6 item*)			
So restless that can't sit still	0.3	11.1	32.00 (3.03, 337.76)
Feel hopeless*	1.0	22.2	21.33 (5.15, 88.39)
So nervous that can't calm down	1.0	11.1	10.67 (1.90, 60.02)
Restless or fidgety*	1.0	11.1	10.67 (1.90, 60.02)
Feel worthless*	1.7	16.7	9.60 (2.48, 37.11)
Feel nervous*	3.8	27.8	7.27 (2.83, 18.72)
So depressed nothing can cheer up*	2.4	16.7	6.86 (1.93, 24.37)
Feel depressed	5.2	27.8	5.71 (2.31, 14.12)
Everything an effort*	1.4	5.6	4.00 (0.47, 34.08)
Tired out	5.2	11.1	2.29 (0.56, 9.32)

Table 4: Convergent validity of depression screening scales with disability and health service utilisation

	PHQ-9	PHQ-2	SRQ-20	Kessler-10	Kessler-6
Disability	**Spearman's Rho correlation coefficient for association with depression screening scales***				
WHODAS 2.0 total score	0.77	0.66	0.74	0.74	0.71
WHODAS 2.0 number of disability days in preceding month	0.60	0.57	0.56	0.58	0.55
Previous health centre visits	**Median depression scale scores (25th and 75th centiles) for number of previous health centre visits**				
0 visits	2 (0, 4)	0 (0, 1)	2 (1, 7)	12 (10, 17)	6 (6, 10)
1 visit	3 (1, 7)	0 (0, 2)	5 (1, 11)	14 (10, 19)	8 (6, 12)
2 or more visits	3 (1, 8)	1 (0, 2)	5.5 (1, 12)	13.5 (10, 20)	8 (6, 12)
χ^2(degrees of freedom), p-value†	7.34 (2) p = 0.0254	2.53 (2) p = 0.2826	6.30 (2) p = 0.0428	1.49 (2) p = 0.4745	2.86 (2) p = 0.2388

PHQ-9 and PHQ-2: Patient Health Questionnaire, 9 and 2 item versions; SRQ-20: Self Reporting Questionnaire, 20 item version; Kessler-10 and Kessler-6: Kessler 10 and 6 item version

*all correlations significant at the level of $p<0.0001$

†Kruskal-Wallis equality of populations rank test

Table 5: Multivariable model for factors associated with misclassification of depression status

	Adjusted odds ratios for association with misclassification of major depressive disorder (MDD) at optimal screening scale cut-off scores[±]		
	PHQ-9 ≥ 5 (n=302)	**SRQ-20 ≥ 8 (n=302)**	**Kessler-10 ≥ 18 (n=302)**
Age (years)	0.99 (0.96, 1.02)	1.00 (0.98, 1.03)	0.98 (0.95, 1.01)
Female sex	0.72 (0.37, 1.39)	0.82 (0.43, 1.56)	0.95 (0.50, 1.80)
Years of education	**0.90 (0.83, 0.97)**	**0.92 (0.85, 0.99)**	**0.92 (0.85, 0.99)**
Rural residence	0.80 (0.40, 1.61)	0.97 (0.50, 1.90)	0.81 (0.41, 1.60)
Standing a long time	**3.01 (1.49, 6.09)**	**2.76 (1.38, 5.55)**	**2.96 (1.49, 5.89)**
Walking long distances	**2.51 (1.13, 5.55)**	**4.85 (2.17, 10.85)**	2.17 (0.99, 4.77)

[±]Gold standard diagnosis of MDD included in the model
PHQ-9: Patient Health Questionnaire, 9 item version
SRQ-20: Self-Reporting Questionnaire, 20-item version

References

1. Wang, P.S., et al., *Delay and failure in treatment seeking after first onset of mental health disorders in the World Health Organization's World Mental Health Survey Initiative.* World Psychiatry, 2007. **6**(3): p. 177-185.
2. Mogga, S., et al., *Outcome of major depression in Ethiopia: population-based study.* British Journal of Psychiatry, 2006. **189**: p. 241-6.
3. Prince, M., et al., *Global mental health 1: no health without mental health.* The Lancet, 2007. **370**: p. 859-77.
4. Dua, T., et al., *Evidence-Based Guidelines for Mental, Neurological, and Substance Use Disorders in Low- and Middle-Income Countries: Summary of WHO Recommendations.* PLoS Med, 2011. **8**(11): p. e1001122.
5. Patel, V., et al., *Global Mental Health 3: Treatment and prevention of mental disorders in low-income and middle-income countries.* Lancet, 2007. **370**: p. 991-1005.
6. World Health Organization, *Mental Health Atlas*, 2011, WHO: Geneva.
7. World Health Organization, *Mental Health Gap Action Programme (mhGAP): Scaling up care for mental, neurological, and substance use disorders*, 2008, WHO: Geneva.
8. World Health Organization, *Mental Health Gap Action Programme Intervention Guide (mhGAP-IG) for mental, neurological and substance use disorders in non-specialized health settings*, 2010, WHO: Geneva.
9. Mitchell, A.J., A. Vaze, and S. Rao, *Clinical diagnosis of depression in primary care: a meta-analysis.* Lancet, 2009. **374**: p. 609-619.
10. Tylee, A. and P. Gandhi, *The importance of somatic symptoms in depression in primary care.* Prim Care Companion J Clin Psychiatry, 2005. **7**(4): p. 167-76.
11. Goncalves, D.A., et al., *Evaluation of a mental health training intervention for multidisciplinary teams in primary care in Brazil: a pre- and posttest study.* General Hospital Psychiatry, 2013. **35**(3): p. 304-308.
12. Gilbody, S., A.O. House, and T.A. Sheldon, *Screening and case finding instruments for depression.* Cochrane Database of Systematic Reviews, 2005(4): p. CD002792.
13. Gilbody, S., T. Sheldon, and A. House, *Screening and case-finding instruments for depression: a meta-analysis.* CMAJ Canadian Medical Association Journal, 2008. **178**(8): p. 997-1003.
14. Lund , C., et al., *PRIME: A programme to reduce the treatment gap for mental disorders in five low- and middle-income countries.* PLoS Med, 2012. **9**(12): p. e1001359.
15. Beusenberg, M. and J. Orley, *A user's guide to the Self-Reporting Questionnaire (SRQ)*. 1994, Geneva: World Health Organisation.
16. Hanlon, C., et al., *Detecting perinatal common mental disorders in Ethiopia: validation of the self-reporting questionnaire and Edinburgh Postnatal Depression Scale.* Journal of Affective Disorders, 2008. **108**: p. 251-262.
17. Tesfaye, M., et al., *Detecting postnatal common mental disorders in Addis Ababa, Ethiopia: validation of the Edinburgh Postnatal Depression Scale and Kessler Scales.* Journal of Affective Disorders, 2010. **122**(1-2): p. 102-108.
18. Youngmann, R., et al., *Adapting the SRQ for Ethiopian populations: a culturally sensitive psychiatric screening instrument.* Transcultural Psychiatry, 2008. **45**(4): p. 566-589.

19. Furukawa, T.A., et al., *The performance of the K6 and K10 screening scales for Psychological distress in the Australian National Survey of Mental Health and Wellbeing.* Psychological Medicine, 2003. **33**: p. 357-362.
20. Kroenke, K., R.L. Spitzer, and J.B.W. Williams, *The Patient Health Questionnaire-2. Validity of a Two-Item Depression Screener.* Medical Care, 2003. **41**(11): p. 1284-1292.
21. Kroenke, K. and R.L. Spitzer, *The PHQ-9: A new depression diagnostic and severity measure.* Psychiatric Annals, 2002. **32**(9): p. 509-515.
22. American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (4th Edition) (DSM-IV)*. 1994, Washington, DC: APA.
23. Gelaye, B., et al., *Validity of the patient health questionnaire-9 for depression screening and diagnosis in East Africa.* Psychiatry Research, 2013. **210**: p. 653-661.
24. Kortmann, F. and S. ten Horn, *Comprehension and motivation in responses to a psychiatric screening instrument. Validity of the SRQ in Ethiopia.* British Journal of Psychiatry, 1988. **153**: p. 95-101.
25. Weobong, B., et al., *The comparative validity of screening scales for postnatal common mental disorder in Kintampo, Ghana.* Journal of Affective Disorders, 2007. **doi: 10.1016/j.jad.2008.05.009**.
26. Sheehan, D.V., Y. Lecrubier, and K.H. Sheehan, *The Mini-International Neuropsychiatric Interview (MINI): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD 10.* Journal of Clinical Psychiatry, 1998. **59(Suppl 20)**: p. 22-33.
27. Sheehan, D.V., Y. Lecrubier, and et al., *The validity of the Mini International Neuropsychiatric Interview (MINI) according to the SCID-P and its reliability.* European Psychiatry, 1997. **12**(5): p. 232-241.
28. World Health Organization (WHO), *World Health Organisation Disability Assessment Schedule II. 12 item interviewer-administered version*, 2000: Geneva.
29. Mogga, S., et al., *Outcome ofmajor depression in Ethiopia: Population-based study.* British Journal of Psychiatry, 2006. **189**: p. 241-246.
30. Buderer, N.M., *Statistical methodology: I Incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity.* Acad Emerg Med, 1996. **3**: p. 895–900.
31. Lauritsen, J.M. and M. Bruus, *Epidata (Version 3). A comprehensive tool for validated entry and documentation of data. ,* 2003, The Epidata Association,: Odense, Denmark,.
32. StataCorp LP, *Stata/SE Version 13.1*, 2014: Texas, USA.
33. Cronbach, L.J., *Coefficient alpha and the internal structure of tests.* Psychometrika, 1951. **16**: p. 297-334.
34. Youden, W.J., *Index for rating diagnostic tests.* Cancer, 1950. **3**: p. 32-35.
35. Lumley, T., R. Kronmal, and S. Ma, *Relative risk regression in medical research: Models, contrasts, estimators and algorithms.*, in *UW Biostats Working Paper Series*2006, University of Washington: Washington.
36. Ludermir, A. and G. Lewis, *Investigating the effect of demographic and socioeconomic variables on misclassification by the SRQ-20 compared with a psychiatric interview.* Social Psychiatry and Psychiatric Epidemiology, 2005. **40**: p. 36-41.

37. Weobong, B., et al., *The comparative validity of screening scales for postnatal common mental disorder in Kintampo, Ghana.* J Affect Disord, 2009. **113**(1-2): p. 109-117.
38. Patel, V., et al., *Detecting common mental disorders in primary care in India: a comparison of five screening questionnaires.* Psychological Medicine, 2008. **38**(2): p. 221-8.
39. Hanlon, C., et al., *Detecting perinatal common mental disorders in Ethiopia: validation of the self-reporting questionnaire and Edinburgh Postnatal Depression Scale.* J Affect Disord, 2008. **108**(3): p. 251-62.
40. American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition) (DSM-V).* 2013, Washington, DC: American Psychiatric Publishing,.
41. Fava, M., et al., *The Importance of Irritability as a Symptom of Major Depressive Disorder: Results from the National Comorbidity Survey Replication.* Mol Psychiatry, 2010. **15**(8): p. 856-67.
42. Manea, L., S. Gilbody, and D. McMillan, *Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis.* CMAJ, 2012. **184**(3): p. E191-6.
43. Pence, B.W., et al., *Validity of an interviewer-administered patient health questionnaire-9 to screen for depression in HIV-infected patients in Cameroon.* J Affect Disord, 2012. **143**(1-3): p. 208-13.
44. Okulate, G.T., M.O. Olayinka, and O.B. Jones, *Somatic symptoms in depression: evaluation of their diagnostic weight in an African setting.* Br J Psychiatry, 2004. **184**: p. 422-7.
45. Adewuya, A.O., B.A. Ola, and O.O. Afolabi, *Validity of the patient health questionnaire (PHQ-9) as a screening tool for depression amongst Nigerian university students.* J Affect Disord, 2006. **96**(1-2): p. 89-93.